A STUDY OF SUPERVISED INTRINSIC SPECTRAL ANALYSIS FOR TIMIT PHONE CLASSIFICATION

Reza Sahraeian, Dirk Van Compernolle .

ESAT, KU Leuven, Belgium

{Reza.Sahraeian,Dirk.VanCompernolle}@esat.kuleuven.be

ABSTRACT

Intrinsic Spectral Analysis (ISA) has been formulated within a manifold learning setting allowing natural extensions to out-of-sample data together with feature reduction in a learning framework. In this paper, we propose two approaches to improve the performance of supervised ISA, and then we examine the effect of applying Linear Discriminant technique in the intrinsic subspace compared with the extrinsic one. In the interest of reducing complexity, we propose a preprocessing operation to find a small subset of data points being well representative of the manifold structure; this is accomplished by maximizing the quadratic Renyi entropy. Furthermore, we use class based graphs which not only simplify our problem but also can be helpful in a classification task. Experimental results for phone classification task on TIMIT dataset showed that ISA features improve the performance compared with traditional features, and supervised discriminant techniques outperform in the ISA subspace compared to conventional feature spaces.

Index Terms— Phone Classification, Manifold Learning, Intrinsic Spectral Analysis

1. INTRODUCTION

The first step in an Automatic Speech Recognition system typically consists in the computation of low dimensional feature vectors from short overlapping segments of speech. After including temporal dynamics, this may e.g. result in the popular 39-dimensional MFCC based feature vector. Even in this 39-dimensional space consecutive frames exhibit great correlation, hence segmental methods using stacks of those frames - and hence using an implicit feature vector in more than 100-dimensional space is not uncommon. Such high dimension inevitably invokes the curse of dimensionality, especially as we may reasonably assume that speech lives in (or close to) a much lower dimensional embedded manifold [1].

High dimensionality is a curse for any pattern recognition problem, both from a performance and a computational point of view. Thus, dimensionality reduction techniques have played a key role in speech recognition research. Linear techniques such as LDA (Linear Discriminant Analysis) and PCA (Principle Component Analysis) have been two of the most popular dimensionality reduction methods in the speech recognition community for several decades. However, for a wide range of physical signals, including speech, there is a nonlinear mapping from a low-dimensional configuration space to a high-dimensional observation space.

Manifold learning methods have been widely used to learn nonlinear projection maps that recover the underlying configuration space. ISOMAP [2], Locally Linear Embedding (LLE)[3], Laplacian Eigenmaps (LE)[4], Diffusion Maps (DM)[5] and manifold regularization [6] are some examples of nonlinear embedding techniques that may drastically reduce the representational dimensionality while preserving the local structure of data points. This class of algorithms has been widely used in machine learning. However, the validity of the manifold structure assumption is necessary for the success of such techniques.

Manifold learning methods have slowly found their way into the speech recognition community in the past decade. Although the articulatory parameterization of the speech production system, presented many years ago [7] [8], indicates the existence of a low-dimensional manifold for certain classes of speech sounds , it was formalized by A. Jansen using the source-filter model of speech production system in [9], where he also proposed to extend the Laplacian Eigenmaps in the framework of unsupervised manifold regularization which is called Intrinsic Spectral Analysis [10]. ISA not only naturally deals with out-of-sample data, which is a common problem for the typical manifold learning methods, but also provides us with data representation. ISA has been used in an unsupervised [11], semi-supervised [12], and fully supervised [10] manner.

In this paper we focus on supervised Intrinsic Spectral Analysis. In previous work [10], no improvement has been reported using supervised ISA features versus other traditional features, except when combining them with traditional features. Here, we set out to prove that ISA can improve recognition performance by itself by taking a number of considerations into account: 1) Using class based graphs to reduce the complexity and improve the performance. 2) Selecting rep-

This research was supported by the fund for scientific research of Flanders (FWO) under project AMODA GA122.10N.

resentative data for each individual class instead of a random selection to ensure the preservation of a data structure. 3) Investigating the effect of linear discriminant method in the intrinsic subspace compared with the extrinsic one. There is one more aspect in which this paper differs from [10]; the latter dealt with binary weighted graphs, while in this paper, results for gaussian similarity weighted ones are also reported.

The remainder of this paper is structured as follows: In section 2 we briefly review the theoretical background of Intrinsic Spectral Analysis. Section 3 introduces the proposed methods. In Section 4 we present experimental results on a phone classification task. Finally we have a discussion and concluding remarks.

2. INTRINSIC SPECTRAL ANALYSIS

Considering a manifold $\mathcal{M}$ embedded in $\mathcal{R}^H$ and a collection of n samples $X = [x_1, x_2, ..., x_n] \subset \mathcal{M}$ that forms a mesh of data points that lie on the manifold, as is typical in manifold learning algorithms, an undirected adjacency weighted (or binary) graph $G = (X, \mathbf{W})$ is constructed with one vertex per data point and the similarity matrix $\mathbf{W} \in \mathcal{R}^{n \times n}$. w_{ij} (the ijth element of $\mathbf{W}$) represents the similarity between x_i and x_j if x_i is one of the κ nearest neighbors of x_j (or vice versa) and 0 otherwise. Then, the so-called graph Laplacian is defined, $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is the diagonal vertex degree matrix with elements $D_{ii} = \sum_{j=1}^{n} w_{ij}$. One can also consider a normalized variant, $\mathbf{L}_{norm} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, where $\mathbf{I}$ is the identity matrix. This normalization reduces the effect of large variation in vertex degree arising from sampling sparsity [13].

Conventional Laplacian Eigenmaps regard the graph as a mesh on the manifold and find the basis determined by the graph Laplacian as an approximation to an intrinsic basis for the manifold that the sample was drawn from [4]. However, this method is limited to the eigen functions of the graph and not the entire manifold. Thus, we seek for a projection f to an intrinsic basis on the manifold. In Intrinsic Spectral Analysis out-of-sample data is approximated by learning such a function in the framework of unsupervised manifold regularization:

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \|f\|_K^2 + \xi \mathbf{f}^T \mathbf{L} \mathbf{f} \tag{1}$$

Where $\mathcal{H}_K$ is the Reproducing Kernel Hilbert Space (RKHS) for some positive semi-definite $n \times n$ kernel function K, $\mathbf{f} = [f(x_1), f(x_2), ..., f(x_n)]^T$ is the vector of values of f for the training data, and $\mathbf{L}$ is the graph Laplacian. ξ is the parameter which makes the balance between extrinsic and intrinsic smoothness of the functions. The lth component of the solution to this optimization problem, based on the RKHS representer theorem, can be expressed as

$$f_l^*(v) = \sum_{i=1}^{n} a_i^l K(x_i, v) \tag{2}$$

$a^l \in \mathcal{R}^n$ is the lth eigenvector (sorted by eigenvalue) to the following generalized eigenvalue problem

$$(\mathbf{I} + \xi \mathbf{L} K)a = \lambda K a \tag{3}$$

In this paper we always use a Radial Basis Function (RBF) kernel: $K(y, x) = exp(-\|y - x\|^2 / 2\sigma^2)$.

3. METHODS AND ALGORITHMS

In [11] supervised phone recognition on the TIMIT dataset using ISA features has been investigated and compared against traditional features such as MFCCs and MLPs; no improvement has been reported. However, the intrinsic coordinates were learnt globally with no labeling information and using binary graphs. Moreover, data points were selected randomly among the whole corpus which is not, indeed, promising to exploit the underlying manifold for each individual class of data. In this section, we propose approaches to deal with these issues.

3.1. Class based ISA

Suppose $c_i \in \{1, 2, .., C\}$ is the label corresponding to x_i, where C is the total number of classes. In this study, the similarity term, w_{ij}, is computed only if x_i and x_j have the same label i.e. $c_i = c_j$. This class based graph has been used for a linear interpretation of Laplacian Eigenmaps (locality preserving projections) to reduce the storage and computational requirements [14]. It is also reasonable to argue that restricting the similarity measure for the points within the same class may improve the classification result since we don't insist on keeping samples of different classes close to each other even though they are nearest neighbors within the ambient space. Thus, the similarity matrix takes the block diagonal form:

$$\mathbf{W} = \begin{pmatrix} W_1 & 0 & \dots & 0 \\ 0 & W_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & W_C \end{pmatrix} \tag{4}$$

Where each W_c is the $n_c \times n_c$ dimension matrix whose elements show the similarity between samples labeled as belonging to class c, and n_c is the number of data points in class c such that $n = \sum_{c=1}^{C} n_c$. In this study, we use the gaussian similarity function, $w_{ij} = exp(-\|x_i - x_j\|^2 / 2\tau^2)$, to exploit more structural information at a cost of setting one more parameter: τ.

3.2. Data Selection

For large scale datasets, including speech datasets, the eigenvalue problem in (3) becomes computationally expensive. Depending on the sparsity and topology of the nearest neighbor graph, the complexity is at least quadratic in the number

of graph vertices. Thus, selecting a subset, $\mathcal{D}$, from the full dataset, $\mathcal{D}_{full}$, with much smaller number of data points and well representative of the structure of data is of great interest.

In this paper, we select a subset of m samples for each individual class of data, and then maximize the nonparametric estimation of the quadratic Renyi entropy for each subset using RBF kernel as has been discussed in [15]. This scheme finds the representative subset of m data points $\mathcal{D} \subset \mathcal{D}_{full}$ such that the quadratic Renyi entropy

$$E(\mathcal{D}, \rho) = -\log \int p(x)^2 dx \approx -\log(\frac{1}{m^2} \mathbf{1}_m^T \mathbf{K} \mathbf{1}_m) \quad (5)$$

is maximized. Where $\mathbf{1}_m$ is a vector of m ones and $\mathbf{K}$ is the $m \times m$ RBF kernel matrix with parameter ρ. This criterion can be maximized iteratively in a greedy manner in order to select points that preserve the underlying structure of the data [16]. To accomplish this, we use the following algorithm:

1. Randomly select a subset $\mathcal{D}$ from the full data set $\mathcal{D}_{full}$
2. Compute the quadratic Renyi entropy of $\mathcal{D}$ using (5)
3. Select a data point x^* from $\mathcal{D}$ and select a data point x^{**} from the remaining pool of data $\mathcal{D}_{full} \setminus \mathcal{D}$.
4. Replace x^* with x^{**} and compute the the quadratic entropy of the new subset.
5. If the entropy in step 3 increases with respect to the entropy of $\mathcal{D}$, then x^* and x^{**} are swapped; otherwise they return to their first subsets.
6. Iterate from step 3 ...

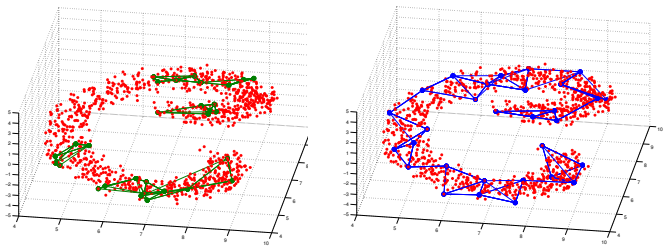
In theory, an appropriate kernel parameter, ρ, also called Parzen window size, corresponds to an appropriate density estimate. We found that the results were not very sensitive to the choice of ρ since we only compare the entropies in each iteration. Silverman's rule [17] is one of the simplest possible choices, given by

$$\rho = \delta[\frac{4}{(2H+1)n_c}]^{[1/(H+4)]} \quad (6)$$

Where H is the dimension of data, δ is the sum of diagonal elements in the covariance matrix of data in $\mathcal{D}_{full}$, and n_c is the number of data points in $\mathcal{D}_{full}$. In our experiments, this algorithm is used for each individual class of data; so, $\mathcal{D}_{full}$ represents a set of data points with same label.

It is worth noting that other clustering and prototype techniques such as k-means or Linde-Buso-Gray [18] may be applicable to find a representative subset as well.

Following is a toy example to visualize how the method works. The full data set is represented in 3-dimensional space with a helix structure. 30 data points are randomly selected to form the $\mathcal{D}$ subset. They are highlighted with rounded green points associating with edges after constructing 3-nearest neighbor graphs (Figure 1.(a)). Using this subset as the initial one and applying the above algorithm, the new subset of data points achieved after 2000 iteration is shown in Figure 1.(b) in blue. It is clear that the new subset is much more representative of the helix structure than the first one.



(a) Random subset (b) Representative subset

Fig. 1. Maximizing the Renyi entropy to select more representative points for the toy dataset with a helix structure.

3.3. Linear Discriminant methods

One of the interesting aspects of intrinsic spectral representation is the improvement in linear separability presented in [10]. A fortunate consequence of this, which is worth investigating, is that the Linear discriminant analyses are expected to be working better in intrinsic subspace than extrinsic one. In this paper we also examine the effects of applying Linear Discriminant Analysis (LDA) [19] to ISA features.

4. EXPERIMENTS

4.1. Experimental Setup

In our experiments on TIMIT, we used the standard NIST training set (462 speakers, 3696 utterances) for training, development set (50 speakers, 400 utterances) in line with [20] to tune the parameters, and the standard core test set (24 speakers, 192 utterances) for testing purpose. 51 TIMIT phone labels are used for training and they are further mapped into the commonly used 39 classes in evaluation phase to calculate the classification accuracy. For feature extraction, a short-time Fourier analysis is performed with a 30ms Hamming window and a 5ms window shift followed by Vocal tract length normalization and mean normalization. Each frame was represented by a 24-dimensional Mel-Spectrum applying triangular shaped filterbank using the full spectrum (24 channels for 16 kHz).

To train the intrinsic coordinates, all features were normalized to have zero mean and unit variance. Then, 300 samples were selected from each individual class of training data as explained in section 3.2. Then, the weighted similarity graph is constructed for each class and total similarity graph is made as explained in section 3.1. to make the normalized graph Laplacian. After finding the intrinsic coordinates by ISA, we kept only the first 13 ones (skipping the first trivial

Table 1. Classification Accuracy on validation set and test set using optimized parameters for the gaussian weighted Laplacian graph

	Validation set	Test set
Accuracy	72.03(%)	70.6(%)

one) and adding the first and second derivatives (Δ and $\Delta\Delta$) features.

4.2. Phone Segmentation and Classification

To describe each phonetic segment in a fixed size feature vector for phone classification experiments, we used the following scenario:

- Step 1: Each phonetic segments is partitioned in 3 sub-segments along the time axis at a 3:4:3 ratio.
- Step 2: Add two more sub-segment from the preceeding and succeeding phones each containing three frames.
- Step 3: Take the average of the local features for each feature and in each sub-segment, and then stack them to a $(5 \times q)$-dimensional supervector representing the phone sample. Where q is the dimension of the speech frame.

In this scenario, the phones with lengths of smaller than 3 frames were ignored. The resulting $(5 \times q)$-dimensional feature vectors form the input for the LDA dimensionality reduction techniques and the classifier. For the phone classification task, a weighted $\mathcal{K}$-Nearest Neighbor classifier [21] is used, where weights are the inverse of the Euclidean distance. For the sake of comparison, we also present results using 13-dimensional Mel-cesptra.

4.3. Experimental Results and Analyses

First of all, we need to find proper values for the parameters; thus, starting from 24-dimensional Mel-spectrum all parameters are jointly optimized on the development set. The resulting optimized parameters are used for the evaluation on the core test set in the rest of this section. The suitable parameters are determined as follows: $\kappa = 30$, $\sigma = 30$, $\xi = 1$, $\tau = 0.5$, $\mathcal{K} = 20$. Table 1. shows the classification accuracies these optimized parameters yield on both validation set and core test set for 13-dimensional ISA features.

Next, velocity (Δ) and acceleration ($\Delta\Delta$) features were included to form 39-dimensional features. 13-dimensional MelSpectrum in Table 2. is obtained by applying PCA to 24-dimensional Mel-spectrum introduced in 4.1. This table shows that (Δ) and ($\Delta\Delta$) features have more effect on extrinsic features than ISA. Due to the closeness in performance, one might reasonably ask how much ISA is disadvantageous

Table 2. Classification Accuracy for MFCC, Mel-Spectrum and ISA features together with their derivatives

Feature set	Accuracy	# of dimension
ISA	70.6(%)	13
MFCC	68.16(%)	13
Mel-Spec	67.13(%)	13
ISA+Δ+$\Delta\Delta$	72.25(%)	39
MFCC +Δ+$\Delta\Delta$	72.16(%)	39
Mel-Spec+Δ+$\Delta\Delta$	71.15(%)	39

Table 3. Classification Accuracy applying LDA to different feature sets

Features	ISA(BW)	ISA(GW)	MFCC	Mel-Spec
Accuracy	75.09(%)	**75.71**(%)	74.2(%)	74.7(%)

in terms of computational complexity comparing to MFCCs. To answer this question we should note that ISA provides us with an intrinsic subspace where linear separability increases while MFCC does not.

In the next part of our experiments, we have investigated the linear separability of intrinsic subspace. As mentioned before, the phonetic separability increases by using intrinsic coordinates even in the unsupervised manner because of the meaningful connection between them and distinctive features of speech sounds [10]. Therefore, LDA approaches are expected to outperform in this subspace. Here, we applied LDA after phonetic segmentation explained in 4.2. for different feature sets. LDA transformation matrix was trained over 51 training labels and it mapped the samples to a 50-dimensional space. It is also interesting to see how using the gaussian similarity weights (GW) instead of the binary weights (BW) in graph construction affects the classification results. To this end, we have conducted the same experiments with binary weighted graphs. Tuning the parameters using validation set yields the same values as mentioned before, except that in the latter there is no τ anymore.

It is important to note that the difference between the accuracies reported in this paper for classification and those in [11] for recognition, e.g. 74.2% versus 76.8% for MFCCs, can be due to the fact that we used the simple $\mathcal{K}$NN classifier and phonetic segmentation in the former while a state-of-the-art hidden markov model/multilayer perceptron back-end was used in the latter to evaluate the recognition. Nonetheless, what we explored in this article, was a comparison between intrinsic and extrinsic subspace in a supervised framework.

5. DISCUSSION AND CONCLUSION

Conventional Intrinsic Spectral Analysis (ISA) is an unsupervised technique. In the supervised approach, however, we may plug in the labeling information to improve the performance. In this paper, we presented our idea to use the labeling

information by constructing the class based graphs. We also proposed to maximize the quadratic Renyi entropy to find a proper subset of data points for each individual class without losing much information regarding the structure of the data in order to deal with the complexity issue by reducing the size of Laplacian graph. Although we have used this data selection method in the ISA framework, it is applicable as a preprocessing box before any manifold learning method. We also plan to investigate the comparison of various data selection methods. Besides, it was shown that the higher linear separability in the intrinsic subspace compared with the extrinsic one leads to higher accuracy using Linear Discriminant Analysis.

This method, however, is highly parametric and finding the proper values for them is not easy. Although, it needs to be done only in the training phase, automatic parameter selection is an important goal. Another problem which is associated with all manifold learning techniques and still remains in this work is the effect of noise which can obscure the manifold structure.

6. REFERENCES

[1] M. Nilsson and W. B. Kleijn, "On the estimation of differential entropy from data located on embedded manifolds," *IEEE Transactions on Information Theory,*, vol. 53, no. 7, pp. 2330–2341, 2007.

[2] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[3] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[4] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data represenation," *Neural Computation*, vol. 16, pp. 1373–1396, 2003.

[5] R. R. Coifman B. Nadler, S. Lafon and I. G. Kevrekidis, "Diffusion maps, spectral clustering and reaction coordinates of dynamical systems," *Applied and Computational Harmonic Analysis*, vol. 21, pp. 113–127, 2006.

[6] P. Niyogi M. Belkin and V. Sindhwani, "Manifold regularization: A geometric framework for learning from examples," *Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[7] G. Fant, "Acoustic theory of speech production," Paris, France, 1970.

[8] K. N. Stevens, "Acoustic phonetics," Cambridge, MA, USA: MIT Press, 1998.

[9] A. Jansen and P. Niyogi, "Intrinsic fourier analysis on the manifold of speech sounds," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2006, pp. 241–244.

[10] A. Jansen and P. Niyogi, "Intrinsic spectral analysis," *IEEE Transactions on Signal Processing*, vol. 61, pp. 1698–1710, April 2013.

[11] A. Jansen and P. Niyogi, "Intrinsic spectral analysis for zero and high resource speech recognition," in *Interspeech*, 2012, pp. 878–882.

[12] A. Jansen and P. Niyogi, "Semi-supervised learning of speech sounds," in *Interspeech*, 2007, pp. 86–89.

[13] U. Von Luxburg, M. Belkin, and O. Bousquet, "Consistency of spectral clustering," *The Annals of Statistics*, pp. 555–586, 2008.

[14] Y. Tang and R. Rose, "A study of using locality preserving projections for feature extraction in speech recognition," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 1569–1572.

[15] M. Girolami, "Orthogonal series density estimation and the kernel eigenvalue problem," *Neural Computation*, vol. 14, pp. 669–688, 2002.

[16] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, "Least squares support vector machines," Singapore, 2002.

[17] B. W. Silverman, "Density estimation for statistics and data analysis," Chapman and Hall, London, 1986.

[18] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications,*, vol. 28, no. 1, pp. 84–95, 1980.

[19] K. Fukunaga, *Introduction to statistical pattern recognition*, Access Online via Elsevier, 1990.

[20] A. K. Halberstadt, *Heterogeneous acoustic measurements and multiple classifiers for speech recognition*, Ph.D. thesis, Massachusetts Institute of Technology, 1998.

[21] S. Tan, "Neighbor-weighted k-nearest neighbor for unbalanced text corpus," *Expert Systems with Applications*, vol. 28, no. 4, pp. 667–671, 2005.